基于独立成分分析的射频干扰信号消除方法*

戴伟[1,3,4]，尚振宏[2*]，徐永华[1]，刘辉[2]，杨亚光[2]，强振平[5]

(1.中国科学院云南天文台，云南 昆明 650011; 2. 昆明理工大学信息工程与自动化学院，云南 昆明 650500; 3. 昆明理工大学云南省计算机技术应用重点实验室，云南 昆明 650500; 4. 中国科学院大学 北京 100049; 5. 西南林业大学大数据与智能工程学院，云南 昆明 650224)

摘 要：射电天文已成为人类研究宇宙的重要途径。但随着人类生产、生活的发展，射频干扰信号对射电天文观测的影响越来越严重，观测数据的好坏关系到科学成果的质量甚至结论的真伪。目前广泛采用基于阈值判断射频干扰，对干扰信号直接舍弃部分观测数据的方法。此类方法存在阈值确定困难、观测带宽和时间被缩减等问题。本文针对脉冲星观测射电信号中，各干扰信号及射电信号统计独立以及呈现出的非高斯性，利用独立成分分析对混合信号进行分解；并根据观测信号中脉冲星信号和干扰信号的分布特点，识别脉冲星信号，实现干扰信号消除。使用该方法对云南天文台40米射电望远镜接收到的脉冲星观测信号进行独立成分分析，分解出独立的RFI信号和脉冲星信号，消除射频干扰信号。本文方法在干扰信号消除、射电信号保留及信噪比方面均取得良好效果。

关键词：射频干扰；　独立成分分析；　脉冲星；　干扰信号消除

中图分类号：P111.44　　**文献标识码：**A　　**文章编号：**

0 引 言

射电信号正在成为人类研究宇宙的重要窗口[1]。尤其是针对脉冲星检测和观测已成为射电天文的重要研究内容。从凝聚态到量子色动力学，以及一系列涵盖恒星演化、星际介质、宇宙学的天体物理学主题，脉冲星为其提供了不可替代的实验环境[2,3]。通过脉冲星观测直接证实了太阳系外行星的存在[4]，为暗物质的本质及分布的研究提供了条件，并首次为引力波的存在提供了间接证据[5]。

然而，射频干扰（Radio Frequency Interference，RFI）成为上述研究的重要挑战[6]。在射电天文中，RFI广义上指由人类生产和生活活动而产生的无线电信号，包括电视信号、调频无线电传输、全球定位系统(GPS)、手机和飞机导航通讯等，对接收的微弱天文信号造成的影响[7]。不同来源RFI在时频特性上的差异使得针对所有类型的RFI信号进行建模变得非常复杂和困难[7]。为减少RFI干扰，射电望远镜在选址时，通常在没有或者很少受到RFI影响的地理环境中建造望远镜，并和相关政府部门协调规划出无线电宁静区。但随着射电天文仪器灵敏度不断提升，接收到非天文信号的干扰也越来越明显。宽带、广播和通信中频谱的大量使用，以及越来越多的大规模生产、经济和商业活动，低功率人工宽带信号的使用也变得越来越频繁，这些都会产生射电天文数据中很难消除的干扰信号，使已有RFI问题变得愈发严重[8]。例如，云南天文台利用40米射电望远镜S波段（2150MHz~2450MHz）开展脉冲星观测

*基金项目：国家自然科学基金资助项目(61462052、60277005)，云南省重点研发计划项目(2018IA054)，云南省应用基础研究项目(2017FB001)

收稿日期：；　修订日期：

作者简介：戴伟，男，副教授，博士研究生，研究方向：天文技术与方法；尚振宏（通信作者），男，副教授，研究方向：图像处理。Email：szh@kmust.edu.cn

任务。2006年5月投入运行之初RFI相对较少，但由于该望远镜距离昆明市区较近（距离市中心直线距离8.2公里），且随着城市建设不断发展（距东三环直线距离1.3公里），包括2G、3G、4G手机信号频段和WIFI频段的RFI越来越多，这些干扰信号严重影响了日常射电天文观测。例如，图 1展示了云南天文台40米射电望远镜观测到的脉冲星信号。其中除期望的脉冲星信号外，可清楚观测到RFI。RFI产生的来源和射电望远镜运行的机理决定了几乎所有射电望远镜均面临RFI问题。观测数据的好坏关系到科学成果的质量甚至结论的真伪，开展RFI抑制和消除方法研究对射电天文发展具有重要理论意义与实际应用价值。

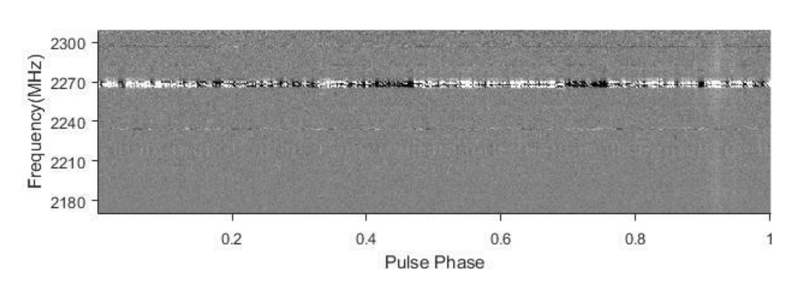


图 1 脉冲星 J0332+5434 频域柱状图

Fig. 1 The phase distributions of J0332+5434 in frequency domain

RFI消除目的在于尽量保持天文信号的前提下，消除射电干扰。依据RFI消除方案在射电天文观测阶段的不同可分为四个环节[8]：观测站预防、预检测、预相关以及应用于干涉数据的后相关方案。对于已经建好投入使用的射电望远镜，后相关处理尤显关键和重要，本文主要针对该环节研究RFI消除方法。

后相关消除方法的基本思想基于：通过统计分析数据来准确标记RFI，即，通过时频二维平面上RFI信号与天文信号形态特性差异标记并消除RFI。天文信号通常呈现宽带、平滑且时间跨度大的特点，而RFI信号经常在时频平面上呈现为高强度像素。目前的RFI消除方法可分为两类。第一类方法采用基于阈值的方法，例如累计和方法[9]与阈值求和方法[6]。这类方法把RFI定义为在时频平面上超过某些阈值的像素。算法简单、高效，被广泛应用于射电数据处理[10]。但此类方法最大的问题在于：如何根据RFI源及观测天体确定阈值？尤其是在针对脉冲星等时变天体信号，阈值选择尤为关键。例如，在对LOFAR的恒星撕裂时间天体Swift J1644+57进行研究时，并未检测到预期的源，分析原因可能是其微弱瞬时信号被认定为RFI而删除[11]。第二类方法采用基于机器学习的方法。近年来，机器学习尤其是深度学习技术在众多领域取得了令人瞩目的研究成果，已有研究将机器学习中的有监督学习以及深度学习的方法应用于RFI消除，取得一定研究进展。例如文献[12]基于K近邻（k-Nearest Neighbor）和混合高斯模型（Gaussian Mixture Models）对RFI信号进行聚类，从而实现RFI标记；文献[13]对基于ANN（Artificial Neural Network）、Adaboost、GBC（Gradient Boosting Classifier）和XGBoost（eXtreme Gradient Boosting）实现的四种有监督学习RFI分类方法的效果进行了分析和比较。但该类基于有监督学习方法的关键问题是：RFI分类准确度对特征选取非常敏感。为了减少对特征的依赖，Akeret等将深度学习的方法应用于RFI消除[1]，对模拟RFI信号取得了非常好的效果。但这种采用模拟数据对深度网络进行训练的方式，很难防止过拟合。

独立成分分析（Independent Component Analysis，ICA）起源于盲源信号分离（Blind Signal Separation，BSS）。BSS是信号处理中一个传统而又极具挑战性的问题，指仅从若干观测的混合信号中恢复出无法直接观测的各个原始信号的过程[14]。这里的“盲”，既指源信号不可测，又指混合系统特性事先未知。所谓“鸡尾酒会问题”就是BSS的典型例子。ICA是研究BSS的一个重要方法，基于信号高阶统计特性已成为阵列信号处理和数据分析的有力工

具。显然，ICA所涉及的问题，在数学模型上本身是欠定的，但附加上原始信号间统计独立及原始信号非高斯分布两个条件后，各原始信号可完美复原[15]。本文中，我们将射电望远镜观测到的脉冲星信号看作观测信号，而包含其中的各RFI和脉冲星信号视为原始信号。各RFI信号和脉冲星信号间统计上相互独立且各信号符合非高斯分布，满足ICA假设条件。相比已有方法，首先，无需人为选择或构造RFI结构特征，不存在阈值选择的困惑；其次，不存在训练或学习过程，因此无需考虑构建训练样本的问题。使用该方法对云南天文台40米射电望远镜接收到的脉冲星观测信号进行独立成分分析，分解出独立的RFI信号和脉冲星信号，进而实现RFI消除，取得良好效果。

1 独立成分分析

ICA是近年来发展起来的一种统计方法。该方法目的是将观察到的数据进行某种线性分解使其分解成统计独立的成分。为严格定义ICA模型，使用统计隐变量模型表示n个同一时刻接收到的射电信号x_1，…，x_n（混合信号）：

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n, i = 1 \cdots n \tag{1}$$

其中，s_i表示包含在混合信号x_i中的源信号（独立成分），即，RFI信号或脉冲星信号。这里混合信号x_i和独立成分s_i均可视为随机变量。不失一般性，设x_i和s_i为零均值（混合信号x_i总是可以通过减去样本均值实现零均值化）。(1)式的矩阵形式为：

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2}$$

其中，**x**和**s**分别为由x_1，…，x_n和s_1，…，s_n组成的随机向量；矩阵**A**由(1)式系数元素a_{ij}组成，称为混合矩阵。(2)式即为ICA模型，描述通过独立成分s_i得到混合信号x_i的过程。目标是求解独立成分**s**，(2)式可调整为：

$$\mathbf{s} = \mathbf{W}\mathbf{x} \quad , \tag{3}$$

其中，矩阵$\mathbf{W}=\mathbf{A}^{-1}$，称为解混矩阵。独立成分**s**为隐随机向量，意味着**s**不能被直接观测到，需要根据仅有信息——随机向量**x**估计出混合矩阵**A**和独立成分**s**。显然，该问题在数学模型上欠定。但在附加两个假设情况时问题变得可解：（1）独立成分s_i间相互统计独立；（2）独立成分s_i拥有非高斯分布。RFI产生自人类活动，而天文射电信号产生自宇宙天体，二者之间相互统计独立，且分布上也不会呈现高斯分布，因此电天文RFI消除满足ICA模型假设条件。

1.1 ICA估计

由概率论中心极限定理可知，多个独立随机变量的混合信号趋近于高斯分布。因此，在ICA模型中，若干个独立成分s_i组成的混合信号x_i比任何独立成分s_i更接近高斯分布。于是可使用分离信号的非高斯性作为分离信号之间独立性的度量。

求解基本ICA问题的通用步骤包括三步[15]：（1）数据(混合信号)的预处理，包括中心化、白化。（2）选择或定义非高斯性(独立性)度量，建立目标函数。该函数取极值时，估计出的独立成分之间非高斯性最大。目标函数代表一种分离准则，根据不同分离准则推导出不同ICA估计算法。（3）用某种最优化方法最大(小)化目标函数，实现ICA估计。依据非高斯性度量方法的不同，ICA估计方法可分为基于峰度（Kurtosis）和负熵（Negentropy）。由于基于峰态的ICA估计方法在实际应用中对边缘样本过于敏感，导致其鲁棒性较差[16]，因此本文采用基于负熵的方法对射电天文中的RFI进行估计。

负熵基于信息论中熵的概念。随机变量的熵可视为其所表示信息的自由度，即，越随机，

熵越大。信息论的一个重要结果为：相同方差时，高斯随机变量熵最大[17]，因此可用熵衡量随机变量的非高斯性。随机变量y的熵定义为：

$$H(y) = -\sum_i P(y = a_i)\log P(y = a_i) \tag{4}$$

其中，a_i是y的可能取值。由(4)式可知，熵为负值。为方便描述随机变量的非高斯性，负熵定义为：

$$J(y) = H(y_{gauss}) - H(y) \tag{5}$$

其中，y_{gauss}为与y具有相同方差的高斯随机变量。(5)式表明：对于高斯随机变量负熵为0，而其他情况则非负。从统计理论出发，负熵是对非高斯性估计的最优方法[18]。但使用(5)式计算负熵时，涉及到估计信号的概率密度函数，在实际应用中这往往非常困难，因此通常采用更简单的方式近似计算负熵。

1.2 负熵的近似

基于最大熵原理，Hyvärinen提出了对负熵的近似方法[19]：

$$J(y) \approx \sum_{i=1}^{p} k_i \left[\mathrm{E}\{G_i(y)\} - \mathrm{E}\{G_i(v)\} \right]^2 \tag{6}$$

其中，k_i为正常数，v为零均值和单位方差的高斯随机变量，G_i为非二次函数。需要注意的是(6)式非负，且当随机变量y呈现高斯分布时，其值为零。

当仅使用一个非二次函数时，对任意非二次函数G，(6)式近似为：

$$J(y) \propto \left[E\{G(y)\} - E\{G(v)\} \right]^2 \tag{7}$$

关键之处在于函数G的选择。一般情况下，选择非快速增长函数G，通过(6)式可得到更鲁棒的负熵估计。例如：下列G函数被证明在负熵估计中非常有效[19]：

$$G_1(u) = \frac{1}{a_1}\log\cosh a_1 u, \qquad G_2(u) = -\exp\left(-u^2/2\right) \tag{8}$$

其中，a_1 为介于 1 与 2 间的常数，通常取 $a_1 = 1$。实际应用中可使用 FastICA 算法[20]寻找(7)式所表示的非高斯性最大值。

2 基于ICA的RFI消除的实现

上面介绍了ICA模型、负熵近似以及快速迭代求解对比函数的方法。为使算法更为高效，将ICA运用到射电天文RFI消除前，需考虑数据初始化问题。此外，如何从ICA分解得到的独立成分中选择出脉冲星信号，也是实现的关键。

2.1 射电数据预处理

实现中，数据预处理包括中心化和白化。

所谓中心化，即，将观测到的射电信号处理为零均值。可通过观测向量**x**减去均值向量**m**=E{**x**}实现。此时，对(2)式两边取期望可知，独立成分**s**也为零均值。中心化的目的在于简化ICA估计。使用中心化后的射电观测数据估计出混合矩阵**A**后，可将**s**的均值向量$\mathbf{A}^{-1}\mathbf{m}$加到

零均值独立成分**s**上，从而完成独立成分估计。

白化可减少ICA估计中的参数。所谓白化，即，将观测到的射电信号**x**转换为非相关且具有单位协方差的新向量 $\tilde{\mathbf{x}}$，使得 $\mathrm{E}\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}}\}=\mathbf{I}$，可通过对**x**协方差矩阵特征值分解实现白化：

$$\mathrm{E}\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}=\mathbf{E}\mathbf{D}\mathbf{E}^{\mathrm{T}} \tag{9}$$

其中，**E** 为 $E\{\mathbf{xx}^{\mathbf{T}}\}$特征向量的正交矩阵，**D** 为其特征值构成的对角矩阵：$\mathbf{D}=\mathrm{diag}(d_1,\ \cdots,\ d_n)$。注意，(9)式的值可通过不同时刻射电观测信号 $\mathbf{x}(1)$，…，$\mathbf{x}(t)$得到，因此，白化可记为：

$$\tilde{\mathbf{x}}=\mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^{\mathrm{T}}\mathbf{x} \tag{10}$$

其中，$\mathbf{D}^{-1/2}=\mathrm{diag}(d_1^{-1/2},\cdots,d_n^{-1/2})$。可以验证：$E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}}\}=\mathbf{I}$。将白化后的观测向量 $\tilde{\mathbf{x}}$ 带入(2)式可得：

$$\tilde{\mathbf{x}}=\mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^{\mathrm{T}}\mathbf{x}\mathbf{A}\mathbf{s}=\tilde{\mathbf{A}}\mathbf{s} \tag{11}$$

其中 $\tilde{\mathbf{A}}=\mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^{\mathrm{T}}\mathbf{x}\mathbf{A}$，为新的混合矩阵。由

$$\mathrm{E}\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathrm{T}}\}=\tilde{\mathbf{A}}\,\mathrm{E}\{\mathbf{s}\mathbf{s}^{\mathrm{T}}\}\tilde{\mathbf{A}}^{\mathrm{T}}=\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\mathrm{T}}=\mathbf{I} \tag{12}$$

可看出，白化后，新的混合矩阵 $\tilde{\mathbf{A}}$ 为正交矩阵，因此只有 $n(n\text{-}1)/2$ 个自由度。若不进行白化，则需估计混合矩阵 **A** 的 n^2 个参数。

为便于表述，下面所描述的射电观测信号**x**为中心化和白化后的数据，而混合矩阵统一表述为**A**。

2.2 主信号分离

Comon Pierre证明了ICA的含混性[21]——ICA不能确定独立成分的顺序，即，通过ICA模型，可由射电观测信号**x**估计出混合矩阵**A**（或解混矩阵**W**）进而得到源信号**s**，但无法确定**s**的组成成分s_1，…，s_n中，哪一个是脉冲星信号，哪些是RFI。

观察图 1所示脉冲星观测数据可发现：（1）数据列在相位上对齐。因此脉冲星观测信号（图像）**x**的每一行可视为一个相变信号：$x_1(t)$，…，$x_n(t)$，由此估计出的源信号向量**s**也由相变信号$s_1(t)$，…，$s_m(t)$组成，为保证**A**是满秩矩阵，要求$n\geq m$；（2）相对于RFI信号，脉冲星信号呈现出强宽带特性，即，对于脉冲星信号，数据行与行之间有更强相关性。此外，由(2)式可知矩阵**A**第i行元素实质上是将源信号$s_1(t)$，…，$s_m(t)$混合成第i个频率通道观测信号$x_i(t)$的权重；第j列元素对应于将第j个源信号$s_j(t)$混合到各频率通道观测信号$x_1(t)$，…，$x_n(t)$的权重。由混合矩阵**A**各元素含义及上述发现（2）可知，若$s_j(t)$对应脉冲星信号，则矩阵**A**第j列将呈现均匀分布的特点，可用方差或标准差衡量随机变量均匀分布程度。此外，实际中RFI源数目不可预知，且一般有n大于RFI源数目，由此会造成矩阵**A**某些列标准差很小且中值也很小，此时这些列对应RFI信号。根据以上分析，可用标准差向量**std**与中值向量**med**点除得到判别向量**d**，即，某一列中值越大标准差越小则对应该判别向量元素值越大：

$$\mathbf{d}=\mathbf{med}\cdot/\mathbf{std} \tag{13}$$

其中，标准差向量和中值向量分别由矩阵**A**各列标准差$std_1, \cdots, std_m$和中值$med_1, \cdots, med_m$组成，即

$$std_j = \frac{1}{N_j}\sum_{i=1}^{n}\sqrt{(a_{ij}-\mu_j)^2},\quad med_j = \left|\mathrm{median}(a_{ij})\right|, i=1,\cdots n, j=1,\cdots m. \tag{14}$$

其中，μ_j 和 N_j，分别表示矩阵 **A** 第 j 列均值和元素个数，median(·)表示中值函数。若第 j_p 个源信号 $s_{j_p}(t)$ 对应脉冲星信号，则其余 $s_j(t)$（$j\neq j_p$）对应 RFI 信号。由上述分析可知，应选择判别向量元素最大值所在列对应的源信号为脉冲星信号。即 j_p 为

$$j_p = \arg\max_j (d_j),\quad j=1,\cdots m \tag{15}$$

其中，$d_1, \cdots, d_m$ 组成判别向量 **d**。需注意，此时 $s_{j_p}(t)$ 包含了各频率通道脉冲星信号信息，并非图 1 展现形式。由混合矩阵 A 各元素含义可知，通过以下变换可得到以图 1 形式呈现的去除 RFI 后脉冲星射电信号：

$$\mathbf{x}_p = \mathbf{A}_p s \tag{16}$$

其中，矩阵$\mathbf{A}_p$通过保留混合矩阵**A**第j_p列元素，并置其余列元素为0得到。

2.3 次信号分离

由于FastICA是对负熵求近似解，少量脉冲星信号会包含在其他独立成分$s_j(t)$（$j\neq j_p$）中。需要对矩阵$\mathbf{A}_p$进一步修正，使得恢复出的信号$\mathbf{x}_p$尽可能完整包含脉冲星信号，同时尽量少包含RFI。注意到残留脉冲星信号不论在频率上还是脉冲相位上分布相对RFI更均匀。通过实验我们发现，可将k倍矩阵**A**除去第j_p列后的标准差作为阈值，并把矩阵**A**中满足

$$\left|a_{ij}\right| < k\times \mathrm{std}(\mathbf{A}_{\bar{p}}) \tag{17}$$

的元素加入$\mathbf{A}_p$中，以修正$\mathbf{A}_p$。修正后的$\mathbf{A}_p$通过式(16)可较完整恢复脉冲星信号。(17)式中 $\mathrm{std}(\mathbf{A}_{\bar{p}})$ 表示矩阵**A**除去第j_p列元素后的标准差。系数k选取过小则少量脉冲星信号会被当作RFI被消除，过大则RFI消除不彻底。但需要指出的是：本文算法对k值并不敏感，取$0.5\le k\le 2.5$可获得信噪比大于50的脉冲星积分轮廓，满足改善脉冲星观测的需求。在遵循消除RFI且尽量保持脉冲星信号这一原则下，我们选取$k=1$。关于k值对RFI消除结果的影响详见本文实验部分。

3 实验

3.1 实验数据

实验数据来自云南天文台40米射电望远镜S波段（2150MHz~2450MHz）脉冲星日常观测结果，观测终端为从澳大利亚CISRO引进的脉冲星观测终端PDFB4（Pulsar Digital FilterBank 4）。PDFB4对观测数据处理过程包括：非相干消色散和周期轮廓折叠。数据采样实现对脉冲星信号数据采集（采样率为64us）。观测中心频率2256MHz，脉冲星的观测方式为非相干消色散，观测配置为512MHz-512Bin-512Chan，30秒的子积分，数据存储格式为PSRFITS格式。由于40米射电望远镜距离昆明市区较近，存在较强RFI，如2G、3G、4G手机信号频段和WIFI频段。为避免引起观测设备系统饱和，观测时需利用滤波器设备在射频

波段直接将上述信号剔除，剔除上述干扰后，S波段仅留下60MHz至140MHz相对干净的观测带宽。实际观测频段为2170 MHz~2310MHz。在该带宽内仍存在各种类型的未引起系统饱和的干扰。

3. 2 实验结果

为验证本文方法的效果和可行性，利用该方法对J0332+5434脉冲星观测数据进行处理。对此我们选取该脉冲星在2017年S波段日常观测中比较有代表性的观测数据。图 2(a)非常清晰地显示出观测干扰频点分别为2231~2241MHz，2252MHz~2258MHz，2267MHz~2273MHz，2300MHz~2308MHz。其中第2、3个干扰频点呈现出干扰信号强、持续时间长的特点，而第1个干扰频点非常弱，通过多个子积分后在图 3(a)中相对明显，第4个干扰频点在相位上显现出不确定性。实验结果如图 2(b)所示。从图中可看出RFI基本消除，仅保留下了PSR J0332+5434脉冲星信号。图 2(c)则显示了图 2(a)与(c)的差值信号，即，FRI信号。

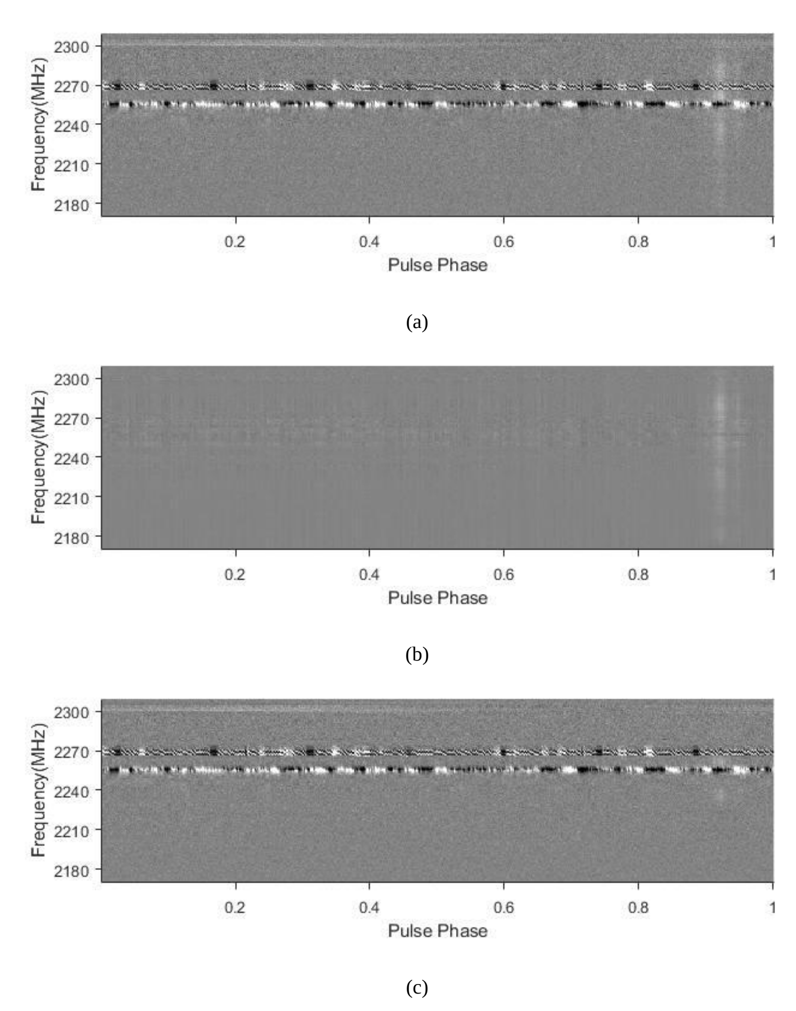


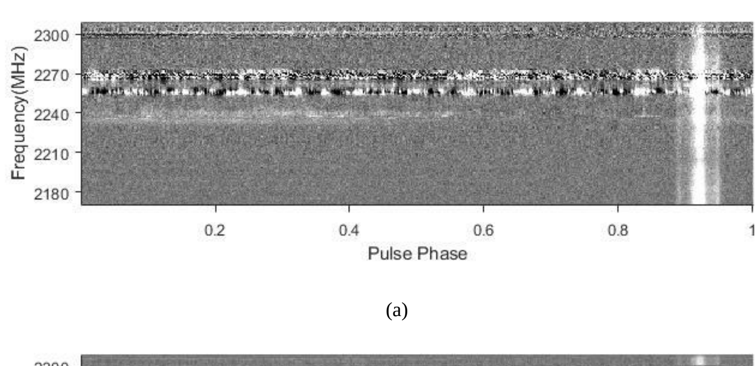
图 2 J0332+5434 RFI 消除效果；(a)观测数据；(b)RFI 消除结果；(c)差值信号

Fig.2 RFI mitigation for J0332+5434 (a) observed signal (b) result of RFI mitigation using ICA (c) difference signal

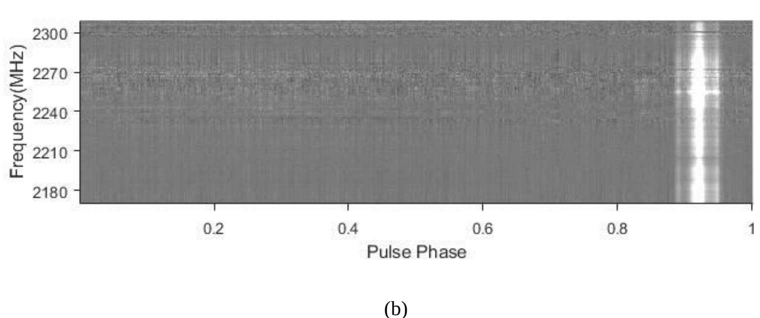
通过此实验，还可对与ICA在原理上有相似之处的独立成分分析方法（Independent Component Analysis，PCA）进行对比分析和讨论。ICA和PCA同属因子分析，都是通过线性组合后使得某种特征最大化。PCA寻求方差最大化，而ICA寻求非高斯性最大化；PCA找出信号中不相关部分，对应二阶统计量分析，ICA找出构成信号的相互独立部分，对应高阶统

计量分析。但PCA和ICA用途不同。PCA是目前数据降维的常用方法，如果只在意数据的能量或方差，假设干扰信号都比较微弱，可用PCA分离出主要信号。但在射电天文RFI消除中，很多情况下强弱干扰混合在一起，例如，图2(a)显示的脉冲星观测信号中，位于2252MHz~2258MHz，2267MHz~2273MHz频点的干扰明显强于脉冲星信号，而2231~2241MHz频点的干扰信号又比脉冲星信号弱很多，2300MHz~2308MHz频点的干扰信号在强度上与脉冲星信号接近。使用PCA，如果以信号强度度量，很难确定哪些主分量对应RFI，哪些对应射电信号。而在某种意义上，ICA更智能——它不在意信号的能量或方差，只看独立性。给定待分解的混合信号经任意线性变换不会影响ICA输出结果，但会严重影响PCA结果。当然如能找到除能量之外的其他特征，并且在此特征维度下使射电信号能成为主分量，而RFI成为次要分量，则可运用PCA分解射电信号和RFI。

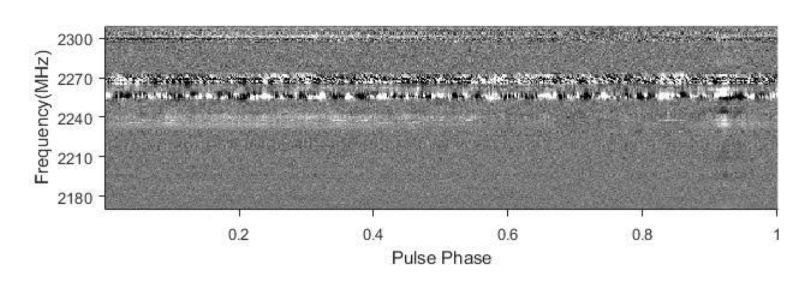
在现实应用中，对脉冲星连续观测信号取均值是一种有效且常用的RFI消手段。图3(a)显示了对PSR J0332+5434脉冲星连续96个子积分观测信号求取均值的结果。图中可明显看到，在脉冲星信号增强的同时，RFI信号也被强化。图3(b)为采用本文方法对96个子积分观测信号分别消除RFI，然后求取均值的结果，图中可看出，在较好保留脉冲星信号同时，RFI信号基本得到消除。图3(c)为图3(a)与(b)的差值信号，即，96个子积分观测信号求取均值后的RFI信号。图3(a)~(c)展示了本文方法对脉冲星观测信号中RFI消除的明显效果。然而仔细观察，图3(c)中出现了少量脉冲星信号残余。这主要由于FastICA是对负熵求近似解，从而导致2231~2241MHz弱RFI信号与残留的少量脉冲星信号对应的独立分量难以区分。但相对于目前普遍采用的均值方法，本文方法明显改善了脉冲星观测信号。这一效果也能从图4所示的脉冲星轮廓中观测得到。



(a)



(b)



(c)

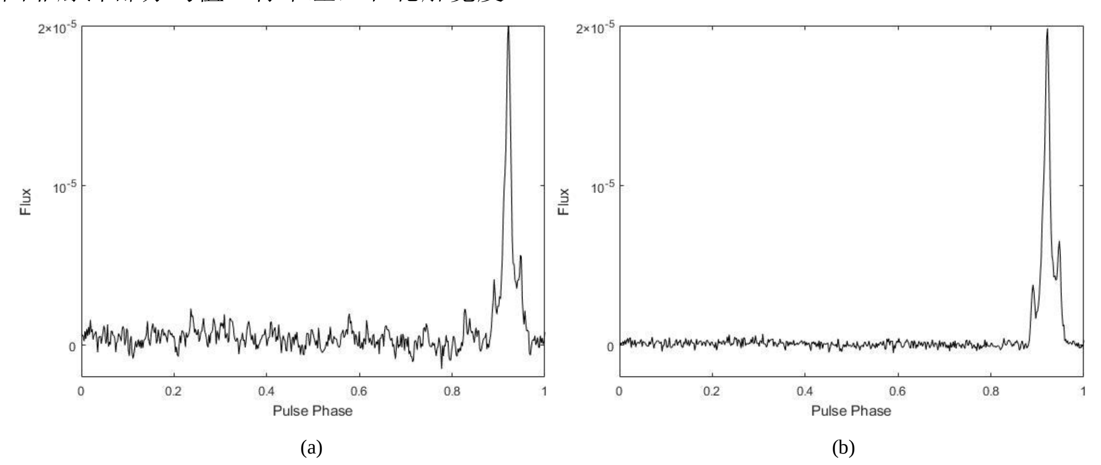
图 3 J0332+5434 积分均值信号 RFI 消除效果：(a) 观测信号积分均值；(b) RFI 消除结果；(c) 差值信号

Fig 3. RFI mitigation for J0332+5434 after mean of accumulation: (a) observed signal after mean of accumulation (b) RFI mitigation result (c) difference signal

为更直观反映本文方法的有效性，图 4显示了对图 3(a)、(b)沿频率域积分后脉冲星轮廓图。其中，图 4(a)为均值法消除RFI后的脉冲星信号积分轮廓图；图 4(b)为采用本文方法，且式(17)中k=1时消除RFI后的脉冲星信号积分轮廓图。两幅轮廓图中均能清晰看到PSR J0332+5434脉冲星的三峰结构，但对比均值法消除结果可看出，本文方法有效消除了观测信号中的RFI信号，且脉冲星信号得到较完整保留。此外，通过脉冲星轮廓信噪比（Signal Noise Ratio，SNR）定义[22]：

$$S/N = \frac{1}{\sigma_{\mathrm{p}}\sqrt{W_{\mathrm{eq}}}}\sum_{i=1}^{n_{\mathrm{bins}}}(p_i - \bar{p}) \quad (18)$$

可计算出图 4(a)、图 4(b)的信噪比分别为17.69和60.28。其中σ_{p}、$\bar{p}$ 和W_{eq}分别为脉冲星轮廓图非脉冲部分均值、标准差，和轮廓宽度。



(a) (b)

图 4 不同 RFI 消除方法消除 RFI 后 J0332+5434 脉冲星轮廓图对比：(a)均值法；(b)本文方法

Fig 4. Comparison of pulse profiles of PSR J0332+5434 after different method of RFI mitigation: (a) mean of accumulation (b) approach proposed by this paper

图 5显示了式(17)中k取0~3时，采用本文方法对PSR J0332+5434脉冲星连续96个子积分观测信号进行RFI消除并求取均值后积分轮廓图SNR变化情况。由图可看出在k = 0.69时，SNR达到峰值60.37，之后开始衰减。分析原因，k较小时，在FRI消除结果中由式(17)引入的残留脉冲星信号对SNR提升产生的贡献大于由此引入的RFI对SNR的影响，从而SNR上升；但随着k的增加，过多引入弱RFI的负面影响将增大，导致SNR下降。值得注意的是，取0.5 ≤ k ≤ 2.5可获得信噪比大于50的脉冲星积分轮廓。通过式(17)在RFI消除结果中尽量包含完整脉冲星信号是一相互平衡的过程，因此实现时，在遵循消除RFI且尽量保持脉冲星信号这一原则下，并未选取SNR取得峰值时的k值，而是选取k=1。相对目前常用的对脉冲星连续观测信号取均值以高SNR的方法，本文方法对SNR提升效果明显。此效果在图 2~图 4实验结果中也有所反映。

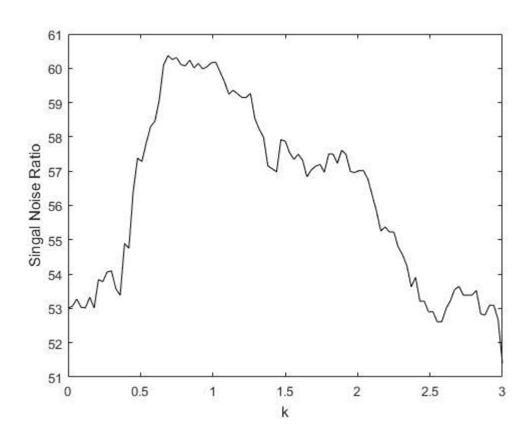


图 5 *k* 值对 RFI 消除信噪比影响

Fig 5. Impact of *k* on signal noise ratio in RFI mitigation

4 结论

RFI广泛存在于射电天文观测中，而脉冲星射电观测信号可视为RFI信号与脉冲星信号的混合信号。基于各RFI信号和脉冲星信号间统计上相互独立且各信号符合非高斯分布的特性，本文提出基于独立成分分析的RFI消除方法。相比已有方法，首先，无需人为选择或构造RFI结构特征，不存在阈值选择的困惑；其次，不存在训练或学习过程，因此无需考虑构建训练样本的问题。使用该方法对云南天文台40米射电望远镜接收到的脉冲星观测信号进行独立成分分析，分解出独立的RFI信号和脉冲星信号，进而实现RFI消除。本文方法在干扰信号消除、射电信号保留及信噪比方面均取得较好效果，为进一步提高脉冲星观测数据利用率奠定良好基础。此外，本文所提方法不仅适用于脉冲星射电信号处理，对其他射电天文信号处理也具有借鉴意义。但是，由于FastICA对负熵求近似解，使得弱RFI信号与残留的少量脉冲星信号对应的独立分量区分度较小，导致出现残差图像中存在少量脉冲星信号的问题。后续研究中我们将进一步深入对该问题的研究。

致谢：感谢国家天文台-阿里云天文大数据联合研究中心对本文工作提供的支持。

Radio frequency mitigation using independent component analysis

Dai Wei[1,3,4], Shang Zhenhong[2*], Xu Yonghua[1], Liu Hui[2], Yang Yaguang[2], Qiang Zhenping[5]

(1. Yunnan Observatories，Chinese Academy of Sciences，Kunming 650011，China; 2. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, china; 3. Yunnan Key Laboratory of Computer Technology Applications, Kunming University of Science and Technology, Kunming 650500, China; 4. University of Chinese Academy of Sciences, Beijing 100049, China; 5. College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming 650224, China)

Abstract: Radio astronomy has become an important way to study the universe. However, with the development of human production and life, radio frequency interference (RFI) has more and more serious impact on radio astronomical observation. The quality of observation is related to the quality of scientific achievements and even the authenticity of conclusions. At present, RFI

detection based on threshold is widely used, and part of the observed data is directly discarded for the interference. Such methods have difficulties in determining threshold values, and reduce observation bandwidth and time. Observed the fact that interferences and radio signals are statistical independence and each is non-Gaussian, we propose a novel approach for radio frequency interference mitigation using independent component analysis to decompose the mixed signal, then identifies the pulsar signal according to the different distribution characteristics between the pulsar signal and RFI signals. The pulsar observations received from 40-meter radio telescope in Yunnan Observatories are processed by the new approach. The results show: RFI signals in pulsar observations are cleanly mitigated while little affecting pulsar signal and good signal-to-noise ratio is achieved.

Key words: Radio frequency interference; Independent component analysis; Pulsar; Interference signal mitigation

参考文献:

[1] Akeret J, Chang C, Lucchi A, et al. Radio frequency interference mitigation using deep convolutional neural networks[J]. Astronomy and Computing, 2017, 18: 35-39.
[2] Blandford R D. Pulsars and physics[J]. Philosophical Transactions of the Royal Society B Biological Sciences, 1992, 341(1660): 177-192.
[3] Ransom S M. Pulsars are cool. Seriously[J]. Proceedings of the International Astronomical Union, 2012, 8(S291): 3-10.
[4] Wolszczan A, Frail D A. A planetary system around the millisecond pulsar PSR1257 + 12[J]. Nature, 1992, 355(6356): 145-147.
[5] Taylor J H. Binary pulsars and relativistic gravity[J]. Reviews of Modern Physics, 1994, 66(3): 711-719.
[6] Offringa A R, De Bruyn A G, Biehl M, et al. Post-correlation radio frequency interference classification methods[J]. Monthly Notices of the Royal Astronomical Society, 2010, 405(1): 155-167.
[7] Fridman P A, Baan W A. RFI mitigation methods in radio astronomy[J]. Astronomy & Astrophysics, 2001, 378(1): 327-344.
[8] 安涛, 陈骁, Mohan, et al. 射电频率干扰的消减[J]. 天文学报, 2017, (5): 18-39.
[9] Baan W A, Fridman P A, Millenaar R P. Radio frequency interference mitigation at the Westerbork Synthesis Radio Telescope: Algorithms, test observations, and system implementation[J]. Astronomical Journal, 2004, 128(2): 933-949.
[10] Akeret J, Seehars S, Chang C, et al. HIDE & SEEK: End-to-end packages to simulate and process radio survey data[J]. Astronomy and Computing, 2017, 18: 8-17.
[11] Cendes Y, Wijers R a M J, Swinbank J D, et al. LOFAR Observations of Swift J1644+57 and Implications for Short-Duration Transients[J]. Physics, 2014.
[12] Wolfaardt C J. Machine learning approach to radio frequency interference (RFI) classification in radio astronomy[J], 2016.
[13] Bethapudi S, Desai S. Separation of pulsar signals from noise using supervised machine learning algorithms[J]. Astronomy & Computing, 2018.

[14] Cardoso J F. Blind signal separation: statistical principles[J]. Proc of IEEE, 2009, 86(10): 2009-2025.
[15] Hyvärinen A, Karhunen J, Oja E. What is Independent Component Analysis?[M]. John Wiley & Sons, Inc., 2003: 145-164.
[16] Huber P J. Projection Pursuit[J]. Annals of Statistics, 1985, 13(2): 435-475.
[17] Cover T M, Thomas J A. Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)[M]. Wiley-Interscience, 2017.
[18] Hyvärinen A, Oja E. Independent component analysis: algorithms and applications[J]. Neural Networks, 2000, 13(4): 411-430.
[19] Hyvärinen A. New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit[J]. Advances in Neural Information Processing Systems, 1997, 10: 273-279.
[20] Hyvarinen A. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis[J]. IEEE Transactions on Neural Networks, 1999, 10(3): 626.
[21] Comon P. Independent component analysis, a new concept?[M]. Elsevier North-Holland, Inc., 1994: 287-314.
[22] Lorimer D R, Kramer M. Handbook of pulsar astronomy[M]. Cambridge university press, 2005.